UNCLASSIFIED

AD NUMBER
ADA175507
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution authorized to DoD only; Specific Authority; AUG 1986. Other requests shall be referred to Marine Corps Headquarters, Code RDS, Washington, DC 20380.**
AUTHORITY
usmc 30 dec 1986

THIS PAGE IS UNCLASSIFIED

CRM 86-189 / August 1986

RESEARCH MEMORANDUM

DETERMINING THE SENSITIVITY OF CAT-ASVAB SCORES TO CHANGES IN ITEM RESPONSE CURVES WITH THE MEDIUM OF ADMINISTRATION

D. R. Divgi

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

3 September 1986

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 86-189

Encl: (1) CNA Research Memorandum 86-189, "Determining the Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves With the Medium of Administration," by D. R. Divgi, Aug 1986

1. Enclosure (1) is forwarded as a matter of possible interest.

2. The Department of Defense may implement a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) in the near future. Analysis of an experimental version has shown that characteristics of items change from paper-pencil to CAT administration. This research memorandum examines the effects of these changes on the CAT-ASVAB scores of individual examinees.

William H. Sims
Director
Marine Corps Manpower
and Training Program

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 86-189

Distribution List

SNDL

A1	ASSTSECNAV MRA	
A1	DASN - MANPOWER (2 copies)	
A6	HQMC MPR	
	Attn:	Deputy Chief of Staff for Manpower (2 copies)
	Attn:	Director, Personnel Procurement Division (2 copies)
	Attn:	Director, Manpower Plans and Policy Division (2 copies)
	Attn:	Director, Personnel Management Division (2 copies)
A6	HQMC TRNG (2 copies)	
A6	HQMC RD&S (2 copies)	
A6	HQMC RA (2 copies)	
A6	HQMC AVN (2 copies)	
E3D1	CNR	
E3D5	NAVPERSRANDCEN	
	Attn:	Director, Manpower and Personnel Laboratory
	Attn:	Technical Library
FF38	USNA	
	Attn:	Nimitz Library
FF42	NAVPGSCOL	
FF44	NAVWARCOL	
FJA1	COMNAVMILPERSCOM	
FJB1	COMNAVCRUITCOM	
FT1	CNET	
V12	CG MCDEC	

OPNAV

OP-01
OP-11
OP-12
OP-13

Other

Joint Service CAT-ASVAB Working Group (15 copies)

CRM 86-189 / August 1986

DETERMINING THE SENSITIVITY OF CAT-ASVAB SCORES TO CHANGES IN ITEM RESPONSE CURVES WITH THE MEDIUM OF ADMINISTRATION

D. R. Divgi

Marine Corps Operations Analysis Group

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Within a few years the Department of Defense may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). In CAT, each test item is characterized by an item response curve (IRC), which describes how the probability of correctly answering the item increases with ability. A recent study conducted by the Center for Naval Analyses found that IRCs of many items in the experimental CAT item pool for the ASVAB changed substantially from paper-pencil to CAT administration. This research memorandum examines the effects of these changes on scores of individual examinees.

EXECUTIVE SUMMARY

INTRODUCTION

Within a few years the Department of Defense may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). Each test question, or item, is characterized by an item response curve (IRC), which describes how the probability of correctly answering the item increases with ability. Existing IRCs for the items intended for use in the Joint Service CAT project are estimated from a paper-pencil (PP) administration.

Using data from an experimental version of CAT, a study conducted at the Center for Naval Analyses (CNA) found evidence that the medium of administration substantially affected the IRCs of many items. This effect is likely to occur with the operational CAT item pool as well; in other words, it is likely that for many items different IRCs will be obtained depending on whether they are calculated from PP or CAT administration. As a result, an item may appear more difficult, or less difficult, in a CAT administration than in a PP administration. This effect creates a potential problem for the CAT-ASVAB project. Strictly speaking, PP-based estimates of IRCs should not be used in CAT; the item pool should be recalibrated, i.e., IRCs should be reestimated using computerized administration. Such recalibration would be very costly in terms of both time and money. The CAT project can proceed without item recalibration only if it can be shown that practical consequences of the medium effect are small enough to be acceptable.

CAT scores have been found to provide as much predictive validity as PP-ASVAB scores. However, this finding does not by itself prove that the medium effect has only a minor impact on CAT scores. One must examine how scores of individual examinees are affected and thus obtain direct evidence to evaluate the impact.

METHOD

The sensitivity of examinees' scores to the medium-of-administration effect was examined using responses of about 7,500 recruits on an experimental version of CAT-ASVAB, contained in the Joint Service Validity Data Set. The data were obtained from the Navy Personnel Research and Development Center (NPRDC).

For each item, the IRC is specified using a mathematical function with three unknown parameters. Item calibration (or recalibration) consists of estimating these parameters for each item. CAT ability estimates derived from PP-based item parameters were compared with those derived from CAT-based item parameters. The PP-based item parameter estimates were provided by NPRDC. They had been obtained using data from a sample of applicants to all the military services. The CAT-based parameters were estimated by the author using *adaptive* tests in the Joint Service Validity Data Set. Estimates were obtained using a simple approximation, without extensive calculations.

Three types of Bayesian ability estimates were analyzed – posterior mode, posterior mean, and the mean of Owen's normal approximation for the posterior distribution. Each estimate was computed twice, using PP-based and CAT-based item parameters. The difference, with the latter subtracted from the former, was the discrepancy in the individual's score due to the medium-of-administration effect.

Mean discrepancy represents the change in mean score for the group as a whole. It has no practical importance because it will be removed when CAT is equated to PP ASVAB. It is the scatter about this mean value that constitutes an additional source of error, that is, variation from one person to another. Standard deviation was used as the measure of the size of this error. Assuming discrepancies to be normally distributed, estimates were made of the percentages of applicants who would have various discrepancies.

RESULTS AND CONCLUSIONS

Considering the simplicity of the estimation procedure, item parameter estimates obtained from adaptive test data were surprisingly good. Over 98 percent of the fitted IRCs passed through the middle of the CAT data points. The chi-square statistics for goodness of fit were at acceptable levels, and any large values occurred because of scatter in the data rather than faulty parameter estimates.

Analysis of discrepancies in individual scores due to the medium-of-administration effect shows that, for all subtests, the error has a spread equivalent to less than one standard score point. Hence even on information subtests, which are sensitive to the medium effect, the size of additional error will only equal or exceed one standard score point for approximately 30 percent of the applicants and two points for about 3 percent of the applicants. Table I shows results by subtest.

All three Bayesian estimators are about equally robust against the medium effect.

Discrepancies become less important when subtests are combined to form composites because errors in different subtests tend to cancel out. Table II shows sizes of discrepancies in Marine Corps composites when the modal estimate of ability is used. The composite with the largest discrepancy is MM, for which approximately 2 percent of applicants are expected to have a discrepancy greater than two composite score points.

Two major conclusions emerge from this study. One is that, given a large enough sample, it is not difficult to estimate item parameters from adaptive test data. The other is that, although IRCs of many items change substantially from PP to CAT administration, the effects on scores tend to cancel out for most examinees. Therefore it appears psychometrically acceptable for the CAT-ASVAB project to proceed without item recalibration based on computerized administration.

TABLE I

ESTIMATED PERCENTAGES OF APPLICANTS HAVING SUBTEST SCORE DISCREPANCIES[a] WITHOUT ITEM RECALIBRATION

	Discrepancies $\geq$ 1 point[b]			Discrepancies $\geq$ 2 points		
	Bayesian estimator:			Bayesian estimator:		
Subtest	Mode	Mean	Owen	Mode	Mean	Owen
GS	2.2	3.3	3.8	<0.1	<0.1	<0.1
AR	1.8	2.4	3.5	<0.1	<0.1	<0.1
WK	1.5	1.8	1.6	<0.1	<0.1	<0.1
PC	12.1	13.4	12.1	0.2	0.3	0.2
AI	21.1	22.2	19.2	1.2	1.4	0.9
SI	29.7	30.4	27.9	3.7	4.0	3.0
MK	4.7	6.7	7.2	<0.1	<0.1	<0.1
MC	28.8	26.1	26.2	3.4	2.5	2.5
EI	15.7	16.5	16.8	0.5	0.6	0.6

[a]Ability estimated from PP-based item parameters minus ability estimated from CAT-based item parameters.

[b]One point is one-tenth of a standard deviation on subtest score scale.

TABLE II

ESTIMATED PERCENTAGES OF APPLICANTS HAVING MARINE CORPS COMPOSITE SCORE DISCREPANCIES[a] WITHOUT ITEM RECALIBRATION

Composite	Discrepancies $\geq$ 2 points[b]	Discrepancies $\geq$ 4 points
CL	<0.1	<0.1
MM	2.2	<0.1
EL	0.1	<0.1
GT	1.5	<0.1

[a]Ability estimated from PP-based item parameters minus ability estimated from CAT-based item parameters.

[b]Two points is one-tenth of a standard deviation on composite score scale.

TABLE OF CONTENTS

INTRODUCTION

Within a few years the Department of Defense may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). Each test question, or item, is characterized by an item response curve (IRC), which describes how the probability of correctly answering the item increases with ability. Each IRC is specified by a mathematical function with three parameters that must be estimated from data. Parameters for the items intended for use in the CAT project have been estimated from a paper-pencil (PP) administration. The issue is whether the parameters need to be reestimated using computerized administration of the items.

Information relevant to this issue is available in data from an experimental version of CAT-ASVAB, known as the Joint Service CAT Validity Data Set [1], provided to the Center for Naval Analyses (CNA) by the Navy Personnel Research and Development Center (NPRDC). Calibration of the experimental items was based on PP administration to a sample of applicants for military service. The experimental CAT system was then administered, using an Apple III computer, to about 7,500 recruits from all four services and a variety of occupational specialties. The data tape provided by NPRDC includes the PP-based item parameters that were used to select items to be administered to each recruit and to estimate the examinee's ability from the responses.

A recent CNA study [2] found evidence that the medium of administration affects item parameters. The study showed that, for many items, the IRC changes substantially from PP to CAT administration. Figure 1 shows an example of such changes. The IRC based on PP item calibration is shown by dots, and the asterisks represent observed proportions of correct answers in CAT administration. It is clear that the item is much harder in the CAT medium of presentation than in the PP medium. The magnitude of the change found in figure 1 is typical of items in the Mechanical Comprehension subtest. The direction of change varies from one item to another. Some items become easier from PP to CAT, while others

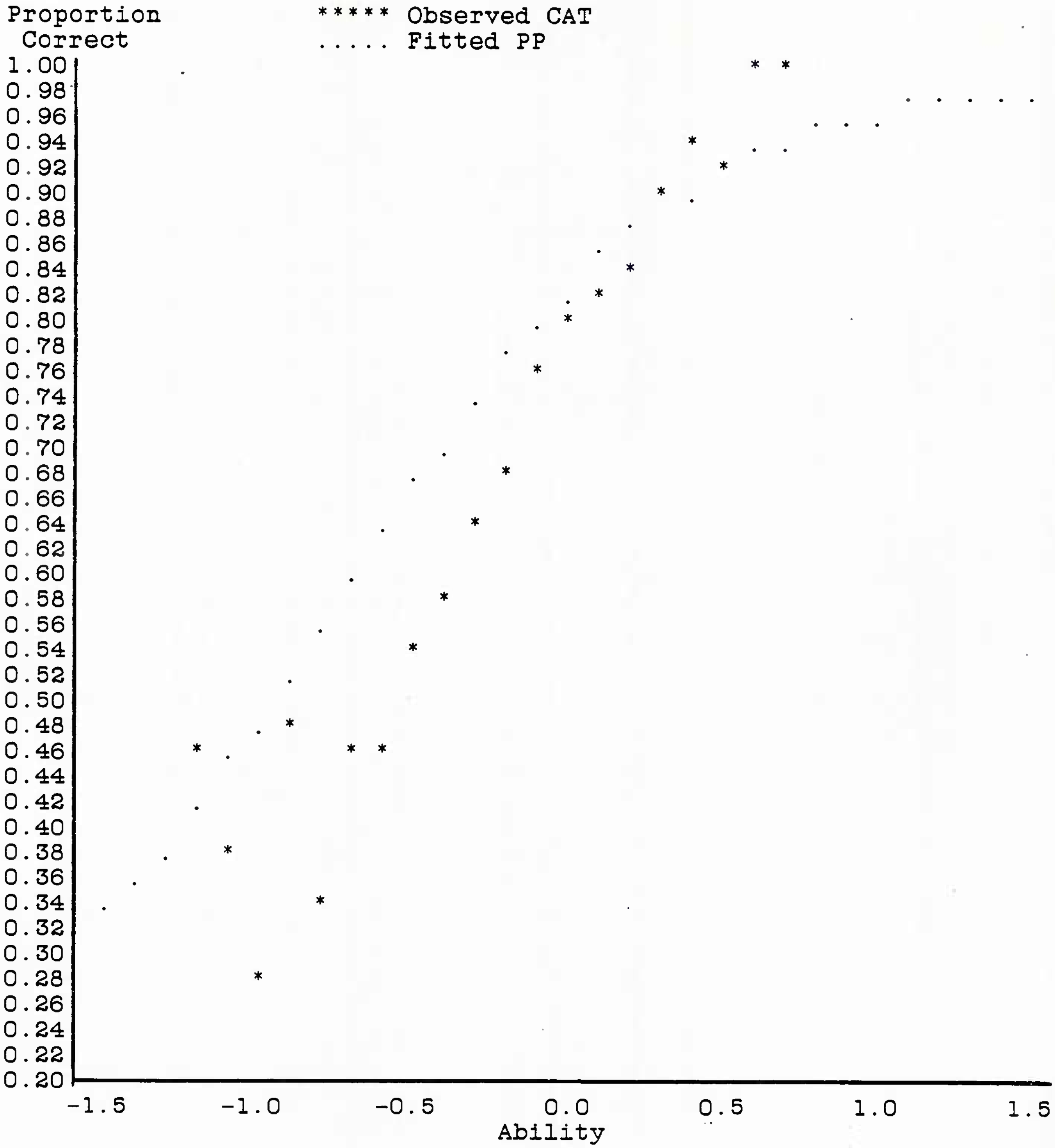

FIG. 1: COMPARISON OF FITTED PP-BASED IRC WITH EMPIRICAL CAT-BASED PROPORTIONS FOR MC ITEM 902, $N = 2{,}556$

become harder. Results reported by Divgi and Stoloff [2] did not reveal any consistent relationship between the change in IRC and the content of the item. Ackerman [3] too, has found that IRCs change substantially from PP to CAT medium of administration and that the changes are not predictable. (Such changes will be called the "medium effect.")

Existence of a medium effect creates a problem for the CAT-ASVAB project. Strictly speaking, PP estimates of item parameters should not be used in CAT; the item pool ought to be recalibrated. In other words, new estimates should be obtained from computerized administration. While this is highly expensive and time consuming, the size of the medium effect on IRCs is too large to be ignored. The project can proceed without item recalibration only if it can be shown that practical consequences of the medium effect are small enough to be acceptable.

The Navy validity study [1] found CAT scores to have as much predictive validity as PP-ASVAB scores. Good validity shows that, despite the existence of a medium effect on IRCs, CAT works well for the group as a whole. However, group-level results are not enough. One must also ask how the scores of individual examinees are affected. If use of PP rather than CAT item parameters leads to a substantial change in the scores of an appreciable fraction of examinees, the use of PP parameters is psychometrically inappropriate.

In the validity study, examinees' scores consisted of Bayesian estimates of ability. Ideally, these should be based on IRCs obtained from computerized administration. If PP item parameters are used instead, the resulting changes in scores constitute a new source of error. Any source of error tends to reduce reliability and validity. Therefore, satisfactory validities suggest that the medium effect causes only small changes in individual scores. On the other hand, it is possible that CAT is considerably superior to PP testing and the two validities appear equal only because CAT validity has been degraded by the medium effect.

Thus, results from the validity study do not yield a clear conclusion. In any case, in view of the importance of the issue, one should not rely on indirect evidence when direct information can be obtained.

Impact on individual scores can be studied using real or simulated data. Segall (attachment 4-5c in [4]) has performed simulations that show how reliability decreases and error variance increases as the size of the medium effect grows. At first sight it appears that his results can be combined with those of Divgi and Stoloff [2] to infer the size of the error introduced by the medium effect. However, this leads to incorrect conclusions because the two studies use different definitions of average deviation. Segall averages the change in IRC over a uniform distribution of ability from -3 to 3 standard deviations. Divgi and Stoloff average it over the examinees who were administered the item adaptively. In the latter case, the distribution of ability tends to be approximately normal and often quite narrow. Therefore, a given value of average deviation represents a much larger medium effect with Segall's definition.

OUTLINE

The analysis consisted of two parts. First, item parameters were reestimated from CAT data using a simple approximation. These were then used to recompute ability estimates for all examinees in the data set. The discrepancy due to the medium effect was obtained by subtracting the new estimate from the original estimate based on PP item parameters.

The distribution of discrepancies was examined separately for each subtest. The nine CAT-ASVAB subtests are General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Auto Information (AI), Shop Information (SI), Math Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI).

ITEM RECALIBRATION USING CAT DATA

Method

Item recalibration, that is, reestimation of item parameters using CAT data, was performed by extending the previous analysis [2]. For each item administered to at least 1,000 persons, Divgi and Stoloff had calculated an ability value for each person who answered that item. It was assumed, purely for convenience, that the distribution of these abilities was normal. Lord and Novick [5] have provided formulas for the proportion of correct answers in the group and for the conditional mean of ability when the item response is correct (equations 16.9.3 and 16.10.3). These expressions are valid when the ability distribution is standard normal and the IRC follows the two-parameter normal ogive model. They are easily modified to the case when mean and variance of ability are different from (0, 1) and the item obeys the three-parameter model.

The three item parameters are discrimination, a, difficulty, b, and guessing, c. In the normal ogive model, the probability that a person with ability θ will answer the item correctly is

$$P(\theta) = c + (1 - c)\ F[a(\theta - b)]\ ,$$

where F is the standard normal, cumulative probability function. Assume that the distribution of ability is normal with mean, μ, and standard deviation, σ. Define

$$b' = (b - \mu)/\sigma$$

and

$$d = \sigma a/(1 + \sigma^2 a^2)^{1/2}\ .$$

Then the proportion of correct answers in the entire group is

$$p = c + (1 - c)\ F(-db')\ ,$$

and the mean ability of examinees who answer the item correctly is given by

$$\mu^+ = \mu + \sigma(1-c)d\ f(db')/p \ ,$$

where f is the standard normal density function.

The CAT-ASVAB project uses the logistic model rather than the normal ogive model for each item, i.e., one in which the normal probability function is replaced by the logistic function [5]. Thus, in the three-parameter logistic model,

$$P(\theta) = c + (1-c)/\{1 + exp[1.7a(b-\theta)]\} \ .$$

On replacing normal probability and density functions with logistic ones in the expressions for p and μ^+, and solving them for item parameters, one obtains

$$\begin{aligned} d &= p(1-c)(\mu^+ - \mu)/1.7\sigma(p-c)(1-p) \ , \\ a &= d/\sigma(1-d^2)^{1/2} \ , \end{aligned}$$

and

$$b = \mu + \sigma\ log\ [(1-p)/(p-c)]/1.7d \ .$$

If the guessing parameter c is assumed known, these equations provide estimates of discrimination and difficulty parameters. The PP estimate of c was used for all items except two (AI item 208 and SI item 505). The exceptions were hard items for which smaller values of c had to be used.

Fit between the estimated IRC and the data points was evaluated by computing the average signed deviation (ASD) and Pearson's chi-square. At each value of ability (rounded to one decimal), the deviation equals the observed proportion of correct answers minus the fitted IRC. ASD is the mean of these differences, with each ability weighted by the number of examinees at that value. It equals the difference between the theoretical and

empirical proportions of correct answers on the item and will be close to zero if the estimation procedure works well.

The chi-square statistic is superior to ASD in that it is sensitive to all differences between the IRC and the data, not merely the proportion of correct answers. It also compares observed deviations with the expected size of random error. On the other hand, a large chi-square cannot be interpreted without a graphical display (as in [2]) to show the nature of the discrepancy.

The medium effect on an item can be quantified in more than one way. Divgi and Stoloff [2] compared the fitted three-parameter IRC from PP administration with observed proportions in CAT administration. Once CAT-based item parameters have been estimated, one can compare the two fitted IRCs. This is somewhat more satisfactory in that the two administrations are treated in the same way. The difference between PP- and CAT-based IRCs was quantified as the average absolute difference (AAD) over all examinees who were administered the item.

No recalibration was performed for items administered to fewer than 1,000 examinees. Such items were found to account for less than 15 percent of the responses in the data set.

Results

Table 1 contains results of item recalibration for the General Science subtest. (Results for the nine subtests are presented in tables A-1 through A-9 in the appendix.) In view of the simplicity of the estimation procedure, the estimates are surprisingly good. Most ASD values are very small, showing that the fitted IRC passes through the middle of the data points. Of a total of 306 items, only 4 show ASD larger than .01 in size: items 208 and 1228 in AI, 1225 and 1526 in SI. Their parameters were refitted by trial and error until ASD fell below .005 in magnitude. The chi-squares fell dramatically for the SI items while remaining almost the same for the AI items.

TABLE 1

RESULTS OF ITEM RECALIBRATION FOR SUBTEST GS

Item	N	ASD*	ChiSq	DF	Item parameters					AAD**
					A		B		C	
					PP	CAT	PP	CAT		
9	4047	0.00	18.4	14	2.16	1.52	0.82	0.71	.14	.07
16	1253	0.00	4.8	4	1.74	1.34	0.52	0.38	.20	.06
126	2433	0.01	32.3	12	2.09	1.65	-0.40	-0.13	.33	.12
202	1602	0.00	6.6	6	1.74	1.57	0.20	0.18	.22	.01
217	4387	0.00	25.7	16	1.89	1.23	0.45	0.30	.16	.06
220	3685	0.00	23.3	20	1.76	1.19	-0.25	-0.66	.20	.07
323	4893	0.00	31.1	17	2.45	1.54	0.27	0.31	.30	.05
326	1523	0.00	13.9	6	1.59	1.46	-0.33	-0.18	.19	.07
423	5578	0.01	31.7	21	2.22	1.65	0.14	0.11	.21	.03
426	3506	0.00	30.8	18	1.85	1.15	0.09	0.00	.26	.04
427	2432	0.00	19.5	14	2.21	1.46	-0.56	-0.65	.29	.03
430	3685	0.01	45.7	19	1.80	1.59	-0.10	0.07	.24	.08
507	1223	0.00	13.9	12	1.60	1.31	-0.82	-1.04	.18	.04
522	1230	0.00	10.1	6	1.70	1.25	-0.53	-0.51	.27	.04
602	1638	0.00	4.9	9	2.49	1.85	1.38	1.39	.13	.04
626	2328	0.00	21.3	8	1.98	1.93	0.91	0.98	.12	.05
720	2843	0.00	20.4	11	2.30	1.64	0.77	0.72	.26	.03
726	3176	0.01	31.9	16	1.96	1.75	-0.43	-0.32	.24	.05
802	2021	0.00	6.5	7	2.11	1.78	0.88	0.98	.18	.07
807	1365	0.00	15.2	6	1.55	0.40	-0.28	-1.92	.18	.11
808	3070	0.01	45.4	18	1.56	1.09	0.05	0.06	.14	.04
814	3042	0.00	10.5	10	1.68	1.45	0.34	0.19	.12	.07
828	3043	0.00	24.4	10	1.59	1.46	0.29	0.38	.10	.05
928	3003	0.00	12.0	9	2.22	2.18	0.88	0.83	.18	.03
929	1655	0.00	9.3	9	2.41	1.99	1.22	1.28	.24	.03
1012	2899	0.00	27.8	13	3.00	3.00	1.10	1.11	.15	.01
1021	1955	0.00	17.5	6	1.80	1.47	0.58	0.58	.20	.01
1119	2147	0.00	8.7	7	1.77	1.42	0.19	0.13	.23	.02
1209	3203	0.00	11.5	16	2.49	1.61	-0.47	-0.75	.33	.03
1212	4120	0.00	25.8	13	2.02	1.68	0.76	0.82	.12	.03
1301	2477	0.00	13.3	8	2.16	2.16	0.75	0.93	.22	.11
1317	1438	0.00	11.0	10	2.56	1.67	1.53	1.49	.10	.07
1324	3162	0.00	19.7	11	2.34	1.74	0.87	0.82	.21	.04
1422	2221	0.00	15.6	11	3.00	2.13	1.32	1.18	.25	.10
1429	1179	0.00	14.7	7	1.94	1.76	1.21	1.36	.13	.09
1515	1254	0.00	4.9	6	1.57	1.57	-0.48	-0.43	.19	.03
1516	2065	0.00	9.4	8	2.09	1.49	0.94	0.92	.17	.03
1523	4581	0.01	56.0	17	1.77	1.56	0.29	0.18	.14	.04

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

Most chi-squares are below 50. (Stoloff and Divgi [2] had suggested that, in view of the large sample sizes, a value below 50 did not indicate presence of a substantial medium effect.) Any large values tend to reflect scatter in the data points rather than incorrect estimation of parameters. This is illustrated in figure 2, which shows observed proportions and fitted IRC for PC item 112, which has the highest chi-square among all items.

The last column of table 1 shows the average change in each IRC from PP to CAT administration, i.e., the size of the medium effect on each item. It varies not only from one item to another but also, on the average, from one subtest to another.

IMPACT ON ABILITY ESTIMATES

Method

For each examinee in an adaptive test, the available information consists of parameters of the items administered and the examinee's responses. There are different ways of scoring these responses, i.e., of computing an estimate of the examinee's ability, θ. The CAT-ASVAB Psychometric Committee has decided that a Bayesian scoring procedure should be used in the CAT-ASVAB project [6]. Bayesian theory begins with a prior distribution of ability that is the same for all examinees and combines it with the individual examinee's item responses to calculate a posterior distribution. (The prior distribution was assumed to be standard normal in this study.) The posterior distribution describes what is known about the examinee's ability. Choice of a scoring procedure consists of choosing a single number to represent the center of this distribution. Three estimators are in general use–the mode and mean of the posterior distribution and the mean of Owen's normal approximation for the posterior distribution [7].

The item parameters are on a scale in which ability estimates have unit variance in the applicant population [1, p. 16]. Standard scores have a

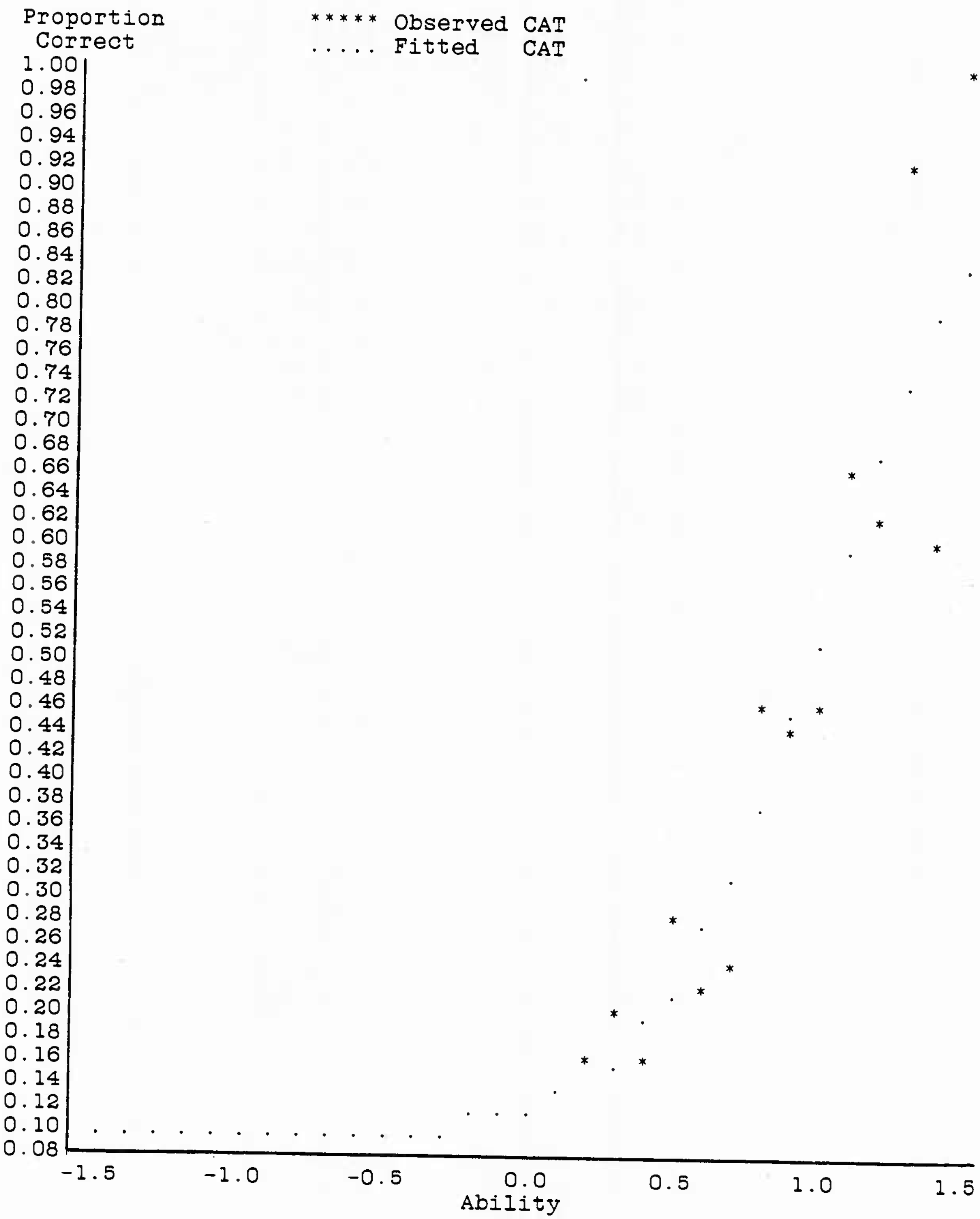

FIG. 2: OBSERVED PROPORTIONS AND FITTED CAT-BASED IRC FOR AN ITEM WITH A LARGE CHI-SQUARE: PC ITEM 112, $N = 2{,}298$

standard deviation of 10 in the national population. Therefore, multiplying any ability estimate by 10 yields a score X which has approximately the same dispersion as standard scores.

Each estimator was computed twice, using PP and CAT estimates of item parameters, and multiplied by 10. Thus, *X(mode, PP)* is the modal score based on PP-based item parameters. For any given examinee, the discrepancy in the modal score is

$$DIS(mode) = X(mode,\ PP) - X(mode,\ CAT) \quad .$$

Discrepancies *DIS(mean)* and *DIS(Owen)* are defined similarly. A positive discrepancy represents benefit and a negative value represents loss to the examinee, resulting from use of PP-based rather than CAT-based item parameters.

For any given estimator, mean discrepancy represents the change in mean score over the entire group. Therefore it has no practical significance. What matters is variation from one person to another, which constitutes an additional source of error. Standard deviation of discrepancy represents the size of this error, and hence the impact of the medium effect on scores of individual examinees.

Results

Table 2 summarizes the results. Mean AAD indicates the average change in IRC from PP to CAT administration. It is clear that the information subtests (AI, SI, MC, and EI) are more sensitive to the medium of administration than the academic subtests.

The standard deviations in table 2 are on the standard score scale. No matter which estimator (mode, mean, or Owen's) is used, error due to the medium effect has a spread of less than one standard score point for all subtests.

To restate the results in more practical terms, table 2 shows percentages of examinees expected to have their scores changed by more than one or two points (assuming discrepancies to have a normal distribution). For four academic subtests, size of the discrepancy exceeds one point for less than 5 percent of examinees if the modal estimator is used. PC is an exception because only 10 items are administered rather than 15. Information subtests AI, SI, MC, and EI exhibit a stronger medium effect and hence larger discrepancies. They are not included in the Armed Forces Qualification Test. In addition, since the operational CAT system will have better graphics than the experimental system, there is reason to expect that the medium effect will be smaller.

Table 2 shows that the three estimators are influenced about equally by the medium effect.

Discrepancies become less important when subtests are combined to form composites because errors in different subtests tend to cancel out. Table 3 shows sizes of discrepancies in Marine Corps composites when the modal estimate of ability is used. The largest is less than one-twentieth of a composite score standard deviation.

CONCLUSIONS

Despite the simplicity of the estimation procedure, results of CAT item calibration were highly satisfactory. This indicates that, given a large enough sample, it is not difficult to estimate item parameters from adaptive test data.

Bayesian estimates of ability are robust against changes in item parameters from PP to CAT administration. Therefore, it appears psychometrically acceptable to proceed without item recalibration, i.e., to use PP estimates of item parameters for computing CAT scores.

TABLE 2

ESTIMATES OF DISCREPANCIES IN SUBTEST SCORES WITH NO ITEM RECALIBRATION*

Sub-test	N	Mean AAD	S.D. of DIS			% |DIS| › 1			% |DIS| › 2		
			Mode	Mean	Owen	Mode	Mean	Owen	Mode	Mean	Owen
GS	7515	.050	.43	.47	.48	2.2	3.3	3.8	‹.1	‹.1	‹.1
AR	6156	.047	.42	.44	.47	1.8	2.4	3.5	‹.1	‹.1	‹.1
WK	7515	.048	.41	.42	.41	1.5	1.8	1.6	‹.1	‹.1	‹.1
PC	6676	.051	.65	.67	.64	12.1	13.4	12.1	0.2	0.3	0.2
AI	6348	.061	.80	.82	.77	21.1	22.2	19.2	1.2	1.4	0.9
SI	6604	.079	.96	.97	.92	29.7	30.4	27.9	3.7	4.0	3.0
MK	6761	.064	.50	.54	.55	4.7	6.7	7.2	‹.1	‹.1	‹.1
MC	6767	.089	.94	.89	.89	28.8	26.1	26.2	3.4	2.5	2.5
EI	6103	.071	.71	.72	.73	15.7	16.5	16.8	0.5	0.6	0.6

* N = Number of examinees
AAD = Average absolute (CAT IRC - PP IRC)
S.D. = Standard deviation
DIS = Discrepancy in ability estimate, on standard score scale
= 10 x (PP based estimate - CAT based estimate)
|DIS| = Absolute value of discrepancy

TABLE 3

ESTIMATES OF DISCREPANCIES IN MARINE CORPS COMPOSITE SCORES WITH NO ITEM RECALIBRATION*

Composite	S.D. of DIS	%|DIS|›2	%|DIS|›4
CL	.48	‹.1	‹.1
MM	.88	2.2	‹.1
EL	.60	0.1	‹.1
GT	.82	1.5	‹.1

* Marine Corps composites have a standard deviation of 20 points. Results for the modal estimates are presented.

REFERENCES

[1] Rehab Group, Inc., *Predictive Utility Evaluation of Adaptive Testing: Results of the Navy Research*, by Susan B. Hardwicke and Kenneth D. White, Dec 1983

[2] CNA, Research Memorandum 86-24, *Effect of the Medium of Administration on ASVAB Item Response Curves*, by D. R. Divgi and Peter H. Stoloff, Apr 1986

[3] Ackerman, T. A., *An Investigation of the Effect of Administering Test Items Via the Computer*, Paper presented at a meeting of the Midwest Educational Research Association, Oct 1985

[4] CNA, Memorandum 86-0454, *Minutes of the February 1986 Meeting of the CAT-ASVAB Psychometric Committee*, by William H. Sims, 17 Mar 1986

[5] Lord, Frederic M. and Novick, Melvin R. *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison Wesley, 1968

[6] CNA, Memorandum 86-0442, *Minutes of the November 1985 Meeting of the CAT-ASVAB Psychometric Committee*, by William H. Sims, 13 Mar 1986

[7] Owen, Roger J. "A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing." *Journal of the American Statistical Association* (June 1975): 351-356

APPENDIX
RESULTS OF ITEM RECALIBRATION

APPENDIX
RESULTS OF ITEM RECALIBRATION

Results of item recalibration for the nine subtests are presented in tables A-1 through A-9.

In each table, the first column contains the code number that identifies each item. The second column shows the number of examinees who were administered that item. This sample size varies from one item to another because items with medium difficulty parameters or high discrimination parameters tend to be used more frequently.

The third column contains the average signed deviation (ASD) between fitted IRC and observed proportions. The ideal value is zero, which occurs when the observed proportion of correct answers, over all persons who were administered that item, equals the value calculated from the fitted IRC.

The fourth column ("ChiSq") presents the chi-square statistic, and its degrees of freedom appear in the fifth column ("DF"). A high chi-square indicates bad fit. In view of the large sample sizes, a chi-square below 50 may be considered to represent satisfactory fit. Values larger than 50 do occur, but infrequently. Examination of plots showed that any large chi-squares resulted from scatter in the data rather than a flaw in parameter estimation.

Both PP and CAT estimates of a and b parameters are shown in the tables. With only two exceptions, the value of c obtained from PP calibration was used in CAT calibration as well.

The last column represents the size of the medium effect, being the average absolute change in the IRC from PP to CAT administration. The average is computed over persons who were administered the item.

TABLE A-1

RESULTS OF ITEM RECALIBRATION FOR SUBTEST GS

Item	N	ASD*	ChiSq	DF	A PP	A CAT	B PP	B CAT	C	AAD**
9	4047	0.00	18.4	14	2.16	1.52	0.82	0.71	.14	.07
16	1253	0.00	4.8	4	1.74	1.34	0.52	0.38	.20	.06
126	2433	0.01	32.3	12	2.09	1.65	-0.40	-0.13	.33	.12
202	1602	0.00	6.6	6	1.74	1.57	0.20	0.18	.22	.01
217	4387	0.00	25.7	16	1.89	1.23	0.45	0.30	.16	.06
220	3685	0.00	23.3	20	1.76	1.19	-0.25	-0.66	.20	.07
323	4893	0.00	31.1	17	2.45	1.54	0.27	0.31	.30	.05
326	1523	0.00	13.9	6	1.59	1.46	-0.33	-0.18	.19	.07
423	5578	0.01	31.7	21	2.22	1.65	0.14	0.11	.21	.03
426	3506	0.00	30.8	18	1.85	1.15	0.09	0.00	.26	.04
427	2432	0.00	19.5	14	2.21	1.46	-0.56	-0.65	.29	.03
430	3685	0.01	45.7	19	1.80	1.59	-0.10	0.07	.24	.08
507	1223	0.00	13.9	12	1.60	1.31	-0.82	-1.04	.18	.04
522	1230	0.00	10.1	6	1.70	1.25	-0.53	-0.51	.27	.04
602	1638	0.00	4.9	9	2.49	1.85	1.38	1.39	.13	.04
626	2328	0.00	21.3	8	1.98	1.93	0.91	0.98	.12	.05
720	2843	0.00	20.4	11	2.30	1.64	0.77	0.72	.26	.03
726	3176	0.01	31.9	16	1.96	1.75	-0.43	-0.32	.24	.05
802	2021	0.00	6.5	7	2.11	1.78	0.88	0.98	.18	.07
807	1365	0.00	15.2	6	1.55	0.40	-0.28	-1.92	.18	.11
808	3070	0.01	45.4	18	1.56	1.09	0.05	0.06	.14	.04
814	3042	0.00	10.5	10	1.68	1.45	0.34	0.19	.12	.07
828	3043	0.00	24.4	10	1.59	1.46	0.29	0.38	.10	.05
928	3003	0.00	12.0	9	2.22	2.18	0.88	0.83	.18	.03
929	1655	0.00	9.3	9	2.41	1.99	1.22	1.28	.24	.03
1012	2899	0.00	27.8	13	3.00	3.00	1.10	1.11	.15	.01
1021	1955	0.00	17.5	6	1.80	1.47	0.58	0.58	.20	.01
1119	2147	0.00	8.7	7	1.77	1.42	0.19	0.13	.23	.02
1209	3203	0.00	11.5	16	2.49	1.61	-0.47	-0.75	.33	.03
1212	4120	0.00	25.8	13	2.02	1.68	0.76	0.82	.12	.03
1301	2477	0.00	13.3	8	2.16	2.16	0.75	0.93	.22	.11
1317	1438	0.00	11.0	10	2.56	1.67	1.53	1.49	.10	.07
1324	3162	0.00	19.7	11	2.34	1.74	0.87	0.82	.21	.04
1422	2221	0.00	15.6	11	3.00	2.13	1.32	1.18	.25	.10
1429	1179	0.00	14.7	7	1.94	1.76	1.21	1.36	.13	.09
1515	1254	0.00	4.9	6	1.57	1.57	-0.48	-0.43	.19	.03
1516	2065	0.00	9.4	8	2.09	1.49	0.94	0.92	.17	.03
1523	4581	0.01	56.0	17	1.77	1.56	0.29	0.18	.14	.04

(Item parameters: A, B, C)

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-2

RESULTS OF ITEM RECALIBRATION FOR SUBTEST AR

Item	N	ASD*	ChiSq	DF	Item parameters A PP	A CAT	B PP	B CAT	C	AAD**
7	3865	0.00	27.8	15	2.41	1.86	0.42	0.30	.17	.06
12	1213	0.00	25.8	13	1.35	0.98	-0.52	-0.79	.10	.06
14	2696	0.00	20.4	9	2.71	1.99	0.74	0.67	.20	.05
19	3777	0.00	36.6	15	2.96	1.80	0.58	0.49	.23	.06
28	3862	0.01	39.9	18	2.11	1.53	0.05	-0.18	.25	.07
301	1374	0.00	25.1	8	2.20	1.72	1.20	1.09	.24	.06
428	2176	0.00	21.0	14	1.55	1.38	-0.41	-0.28	.14	.06
612	3305	0.00	34.8	13	3.00	2.46	0.82	0.84	.16	.03
613	1225	0.00	7.1	4	1.48	1.59	-0.29	-0.09	.14	.10
617	2006	0.00	12.5	6	2.04	1.50	0.10	0.30	.29	.10
709	1429	0.00	6.2	3	2.30	3.00	0.40	0.40	.34	.02
713	3571	0.00	35.9	19	1.90	1.52	0.24	0.42	.14	.09
729	2110	0.00	5.7	10	3.00	2.50	1.05	1.01	.24	.03
801	1970	0.00	10.1	5	2.24	2.50	0.58	0.60	.25	.02
814	1815	0.00	12.3	6	1.89	1.79	0.88	1.00	.10	.08
822	3773	0.01	43.6	19	2.05	1.54	0.08	0.09	.22	.03
902	1186	0.00	5.1	3	1.80	1.98	0.19	0.24	.22	.03
908	2463	0.00	23.1	13	3.00	2.54	1.10	1.12	.17	.02
929	4450	0.01	48.7	18	2.43	2.00	0.23	0.21	.18	.02
1019	1155	0.00	17.0	8	2.03	2.03	1.28	1.26	.18	.01
1021	1152	0.00	12.1	7	2.02	1.61	1.18	1.21	.20	.02
1029	2547	0.00	18.5	9	2.48	1.74	0.92	0.86	.10	.06
1110	2356	0.00	7.4	13	1.60	1.25	-0.23	-0.25	.15	.02
1122	1642	0.00	10.8	7	1.51	1.30	-0.26	-0.28	.12	.01
1205	2942	0.01	27.3	19	1.83	1.22	-0.14	-0.32	.25	.04
1213	3447	0.00	32.0	13	2.52	2.04	0.67	0.70	.12	.03
1227	3399	0.00	19.4	14	2.73	1.77	0.80	0.84	.11	.05
1305	3580	0.00	26.0	14	2.20	1.65	0.55	0.57	.10	.04
1315	1371	0.00	4.0	3	1.69	1.96	0.06	0.16	.19	.06
1410	1355	0.00	10.0	4	1.90	1.40	-0.19	-0.21	.35	.02
1412	1725	0.00	13.1	13	1.57	1.72	-0.66	-0.74	.12	.04
1416	1304	0.00	14.3	10	2.94	1.45	1.62	1.50	.12	.15
1424	2092	0.00	6.4	14	1.74	1.21	-0.53	-0.80	.23	.04
1527	1223	0.00	7.7	7	1.46	1.60	-0.41	-0.34	.13	.03

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-3

RESULTS OF ITEM RECALIBRATION FOR SUBTEST WK

Item	N	ASD*	ChiSq	DF	Item parameters A PP	A CAT	B PP	B CAT	C	AAD**
15	4485	0.01	33.5	17	2.73	1.72	-0.08	-0.11	.27	.04
125	1675	0.00	7.1	5	2.13	1.32	0.23	0.00	.27	.06
313	1745	0.00	3.7	6	2.75	2.48	1.00	0.97	.19	.03
326	4202	0.01	34.0	16	2.22	1.68	0.35	0.38	.15	.04
510	1762	0.00	14.2	9	1.96	2.16	-0.48	-0.46	.17	.01
515	1018	0.00	10.6	8	1.92	1.89	-0.75	-0.67	.30	.03
527	4179	0.00	27.9	18	2.89	1.62	-0.24	-0.57	.28	.05
606	2975	0.00	28.2	12	3.00	2.40	0.84	0.82	.24	.03
711	3729	0.00	15.6	14	2.64	1.50	0.78	0.88	.13	.07
716	3963	0.00	19.5	16	2.83	1.66	0.37	0.39	.33	.05
717	2181	0.00	13.4	6	2.85	2.95	0.90	1.04	.24	.11
718	2637	0.00	23.3	12	3.00	2.46	1.09	1.04	.12	.05
720	1052	0.00	18.5	8	3.00	2.16	1.33	1.28	.20	.05
729	1863	0.00	12.4	5	2.17	1.29	0.26	0.12	.27	.03
806	2715	0.00	10.9	8	1.92	1.57	0.20	0.35	.14	.09
813	2790	0.00	27.7	12	2.50	1.91	-0.31	-0.35	.29	.02
825	4492	0.01	43.9	16	2.71	2.04	0.11	0.21	.34	.06
903	1226	0.00	9.5	6	3.00	3.00	1.25	1.27	.20	.01
912	3733	0.00	33.9	15	3.00	2.46	0.90	0.87	.14	.04
920	3903	0.00	25.9	12	3.00	2.55	0.71	0.65	.24	.04
925	1301	-.01	15.3	10	3.00	2.47	1.36	1.36	.17	.02
1014	1092	0.00	10.8	7	2.69	2.64	1.30	1.27	.14	.02
1020	1330	0.00	4.4	4	2.39	1.67	1.06	1.05	.15	.03
1022	2139	0.00	23.6	10	2.01	1.69	-0.29	-0.37	.19	.02
1120	2201	0.00	9.9	5	2.25	2.53	0.57	0.69	.23	.09
1124	2004	-.01	18.4	11	3.00	3.00	1.20	1.34	.20	.09
1126	2531	0.00	20.2	12	2.45	1.71	-0.36	-0.61	.31	.05
1130	4101	0.01	21.3	11	2.74	2.42	0.53	0.45	.21	.05
1202	2255	0.00	11.3	6	2.60	2.26	0.73	0.76	.25	.02
1213	1881	0.00	16.6	9	1.93	1.51	-0.22	-0.15	.22	.05
1214	5100	0.01	47.2	16	2.80	2.14	0.36	0.39	.21	.04
1311	3997	0.01	41.7	17	2.46	1.52	-0.07	-0.15	.27	.04
1427	4765	0.01	43.8	16	2.31	1.67	0.31	0.13	.14	.08
1521	1319	0.00	10.7	5	2.20	1.33	0.43	0.24	.28	.06
1524	2414	0.00	28.0	9	2.05	1.67	-0.15	-0.06	.21	.06
1526	3392	0.01	49.3	17	2.35	1.57	-0.07	-0.17	.34	.03

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-4

RESULTS OF ITEM RECALIBRATION FOR SUBTEST PC

					Item parameters					
					A		B		C	
Item	N	ASD*	ChiSq	DF	PP	CAT	PP	CAT		AAD**
5	3618	0.00	43.4	12	2.18	1.65	0.64	0.76	.19	.06
7	1455	0.00	18.5	12	1.28	1.96	-0.69	-0.42	.15	.06
8	3489	0.00	27.8	13	2.07	1.74	0.54	0.62	.21	.04
22	5061	0.00	26.9	13	2.64	2.09	0.42	0.46	.20	.04
112	2298	0.00	98.4	11	1.88	2.00	1.11	1.03	.10	.04
118	1123	0.00	15.6	10	1.28	2.01	-0.72	-0.41	.22	.08
128	1192	0.00	7.7	4	1.87	1.26	0.18	0.29	.32	.06
201	2318	0.00	26.4	13	1.52	1.33	-0.24	-0.22	.28	.02
205	1659	0.00	68.4	8	1.77	1.33	0.95	1.00	.11	.04
217	1585	0.00	23.6	8	2.38	2.72	0.75	0.88	.31	.08
218	2845	0.00	25.0	13	1.62	1.53	-0.31	-0.46	.30	.04
224	1895	0.00	27.9	11	1.38	1.23	-0.34	-0.41	.20	.01
226	2503	-.01	22.5	9	3.00	2.46	1.11	1.03	.26	.06
302	2015	0.00	20.5	10	1.61	1.27	0.10	0.06	.26	.01
410	2148	0.00	23.1	9	1.94	0.85	0.63	1.04	.17	.13
424	3799	0.00	30.8	12	2.57	1.64	0.69	0.53	.20	.09
529	4939	0.00	38.9	15	2.09	1.29	0.18	-0.02	.28	.05
530	2183	0.00	22.2	11	2.29	1.52	0.42	0.11	.34	.10
718	1125	0.00	13.3	7	1.42	0.87	-0.04	-0.11	.23	.03
720	2695	0.00	30.7	15	1.98	1.15	0.14	-0.03	.31	.04
806	3612	0.00	36.7	17	1.55	1.15	-0.11	-0.36	.24	.05
1416	2502	0.00	38.2	19	1.61	1.43	-0.29	-0.31	.24	.01
9204	1543	0.01	11.5	14	1.56	1.25	-0.01	-0.11	.26	.02

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-5

RESULTS OF ITEM RECALIBRATION FOR SUBTEST AI

					Item parameters					
					A		B		C	
Item	N	ASD*	ChiSq	DF	PP	CAT	PP	CAT		AAD**
9	1348	0.00	2.9	11	1.18	1.15	1.07	0.91	.21	.06
16	2532	-.01	67.4	18	1.24	2.41	1.15	1.13	.38	.06
18	2762	0.00	19.5	16	1.64	1.00	-0.46	-0.65	.31	.05
116	1822	0.00	22.9	10	1.46	1.75	0.40	0.40	.14	.02
126	1276	0.00	12.5	7	3.00	1.24	-0.58	-0.48	.39	.06
130	3271	0.00	35.2	18	1.50	1.66	0.65	0.54	.11	.04
202	1650	0.00	5.5	7	2.62	2.13	0.32	0.37	.31	.03
204	1229	0.00	12.1	7	2.12	1.87	0.77	0.72	.24	.03
205	1215	0.00	10.8	8	1.22	0.55	-0.11	-0.10	.25	.07
208	1114	-.05	16.0	10	0.18	0.10	2.42	3.00	.15	.04
322	2883	0.00	30.1	18	1.66	1.47	0.60	0.27	.20	.15
418	1730	0.00	13.6	9	1.02	0.92	0.03	0.16	.18	.05
424	1488	0.00	7.6	8	1.14	1.54	-1.04	-0.72	.24	.06
429	1327	0.00	17.6	12	0.65	0.93	2.85	2.42	.12	.01
523	1646	0.00	4.9	8	1.96	2.19	0.71	0.80	.31	.05
623	1817	0.00	10.2	13	1.87	1.37	0.56	0.37	.23	.06
702	1840	0.00	33.4	14	2.21	1.39	1.13	1.39	.35	.07
704	2387	0.00	5.4	10	2.42	1.83	0.23	0.16	.18	.04
721	3108	0.00	30.0	20	1.52	1.27	0.13	0.05	.32	.03
726	3517	0.00	68.1	24	0.65	0.38	-1.47	-0.92	.24	.14
728	2257	-.01	27.0	15	2.15	2.03	0.97	0.91	.20	.04
821	1653	0.00	28.0	10	1.46	1.40	-0.31	-0.06	.39	.09
910	1263	0.00	6.1	9	1.21	1.54	0.83	1.02	.23	.08
912	3919	0.00	57.0	21	1.33	2.25	-0.01	0.32	.37	.09
915	4106	0.01	38.8	23	1.57	1.70	-0.35	-0.10	.29	.08
918	1436	0.00	5.0	8	1.85	1.46	1.41	1.23	.18	.10
924	2478	0.00	11.0	16	1.06	0.97	-0.57	-0.57	.25	.01
1013	1163	0.00	10.5	6	1.43	1.13	0.20	0.46	.18	.04
1026	1197	0.00	10.8	11	0.97	0.51	-1.22	-0.51	.23	.19
1027	3581	0.01	40.8	22	1.52	1.29	-0.24	-0.07	.40	.05
1116	3070	0.01	35.9	21	1.29	1.29	0.03	0.06	.17	.01
1120	2658	0.00	15.4	15	1.67	1.27	0.32	0.16	.13	.05
1122	2468	0.00	54.3	16	1.54	0.66	-0.62	-1.06	.24	.07
1228	1441	0.03	16.9	12	1.09	0.46	-1.27	-3.00	.24	.07
1326	1915	0.00	9.4	12	0.89	0.71	-1.09	-1.22	.24	.01
1402	1541	-.01	36.2	13	1.63	2.03	1.26	1.43	.25	.09
1421	2940	-.01	22.1	17	2.68	2.08	0.95	0.90	.17	.05
1426	1275	0.00	4.2	7	1.61	1.89	-0.05	0.19	.39	.10

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-6

RESULTS OF ITEM RECALIBRATION FOR SUBTEST SI

					Item parameters					
					A		B		C	
Item	N	ASD*	ChiSq	DF	PP	CAT	PP	CAT		AAD**
5	2207	0.00	26.6	17	1.34	1.10	-0.78	-0.96	.29	.05
9	3535	0.00	30.6	19	2.33	1.64	0.05	-0.23	.13	.11
13	1105	0.00	21.1	12	0.82	0.45	-2.24	-2.91	.22	.03
119	2508	0.00	18.9	17	0.98	1.07	-1.85	-0.63	.20	.27
223	2448	0.00	31.4	18	1.16	0.89	0.08	-0.48	.17	.19
224	3399	0.01	40.3	23	1.46	1.16	0.02	-0.25	.18	.10
319	4174	0.00	40.3	27	0.72	0.84	-0.14	-0.29	.16	.05
323	3889	0.00	61.5	23	1.23	0.85	0.74	0.70	.15	.05
328	3147	0.00	22.4	17	1.76	1.22	0.26	0.33	.23	.05
405	2506	0.00	30.0	15	1.61	0.81	0.42	1.00	.23	.17
505	1551	0.00	35.9	12	3.00	3.00	1.19	1.38	.00	.22
524	1806	0.00	26.2	10	1.80	0.92	1.18	0.78	.12	.23
703	2295	0.00	41.0	18	0.73	0.84	1.10	1.15	.17	.02
717	2591	0.00	29.4	12	1.15	1.11	-0.37	-0.23	.20	.05
803	1422	0.00	5.4	9	1.64	1.45	0.18	0.07	.39	.02
817	1062	0.00	22.3	11	2.06	1.25	0.80	0.93	.17	.07
908	1292	0.00	11.5	10	0.96	1.16	-0.79	-0.51	.25	.07
928	2057	0.00	61.4	16	2.28	1.72	1.42	1.46	.22	.03
1013	2419	0.00	27.5	11	0.87	1.01	-0.42	-0.07	.27	.06
1030	4084	0.01	46.6	22	1.97	1.50	-0.35	-0.63	.20	.09
1127	2448	0.00	21.5	15	0.97	1.28	-0.40	-0.33	.26	.02
1204	1132	0.00	27.4	12	3.00	3.00	1.95	1.84	.28	.02
1206	3231	0.00	25.0	18	1.71	1.29	0.17	0.15	.16	.03
1216	1969	0.00	24.5	17	1.08	0.88	0.81	0.80	.21	.02
1225	1511	0.16	189.6	11	0.59	0.10	-2.00	-3.00	.22	.12
1303	2559	0.00	31.0	17	0.26	0.43	1.56	1.12	.22	.02
1319	3626	0.01	34.8	25	1.30	1.06	-0.94	-1.10	.14	.03
1326	1932	0.00	23.6	13	0.77	0.45	-1.69	-2.08	.22	.03
1330	1529	0.00	21.8	18	0.94	1.12	-3.04	-2.60	.20	.01
1409	2099	0.00	8.7	9	1.76	0.99	0.26	-0.04	.28	.09
1411	1005	0.00	11.1	15	0.86	0.81	-2.58	-2.56	.20	.01
1412	4731	0.00	77.5	25	1.51	1.83	-0.49	-0.02	.29	.15
1413	1449	0.00	11.9	7	1.68	1.55	0.00	-0.19	.19	.08
1427	1657	0.00	11.6	13	1.39	0.99	-1.36	-1.25	.22	.07
1428	2114	0.00	30.2	13	1.76	1.34	1.21	1.12	.22	.07
1521	1213	0.00	15.9	11	2.08	1.35	1.62	1.63	.17	.06
1526	2079	0.09	154.3	21	1.24	0.47	-2.51	-3.00	.20	.09

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-7

RESULTS OF ITEM RECALIBRATION FOR SUBTEST MK

Item	N	ASD*	ChiSq	DF	Item parameters A PP	A CAT	B PP	B CAT	C	AAD**
22	2270	0.00	19.2	8	1.80	1.19	0.36	0.33	.13	.04
105	1357	0.00	18.0	7	1.19	0.96	-0.27	-0.48	.13	.06
109	1250	0.00	13.3	5	1.38	0.98	0.20	-0.04	.12	.09
119	2304	0.00	2.4	6	1.97	1.63	0.50	0.67	.15	.09
130	2019	0.00	1.6	6	2.32	1.76	0.97	1.08	.07	.05
215	1675	0.00	5.1	4	3.20	2.50	0.90	1.05	.30	.10
227	1639	0.00	11.9	5	3.33	3.00	1.16	1.06	.14	.10
306	1804	0.00	21.9	9	1.39	1.17	-0.12	-0.17	.15	.02
401	1316	0.00	10.3	4	3.50	2.63	1.28	1.36	.16	.05
402	4641	0.01	90.5	19	2.44	2.08	0.16	0.10	.09	.02
428	3271	0.01	19.7	12	2.19	2.00	0.66	0.52	.11	.09
517	2340	0.00	22.7	11	1.54	1.46	0.10	0.31	.07	.11
518	4067	0.01	34.4	17	2.64	1.71	0.10	0.20	.15	.09
522	1079	0.00	16.7	7	4.00	2.31	1.44	1.38	.22	.07
605	1391	0.00	6.6	4	2.21	2.42	0.95	0.96	.09	.01
619	2007	0.00	24.5	15	1.31	0.91	-0.58	-0.73	.13	.03
625	1399	0.00	0.6	4	2.64	1.63	1.02	1.05	.13	.04
713	3214	0.00	12.6	12	3.98	1.98	0.99	0.85	.16	.12
813	3394	0.00	10.9	12	3.08	1.87	0.71	0.67	.29	.05
826	1523	-.01	12.9	8	3.59	2.66	1.35	1.36	.12	.04
902	4014	0.00	36.1	14	2.27	1.15	0.66	0.45	.07	.10
923	1055	0.00	2.8	3	1.66	1.26	0.47	0.55	.16	.03
1014	1609	0.00	9.2	8	1.34	0.85	-0.26	-1.02	.20	.16
1027	2443	0.00	2.5	6	2.22	1.88	0.64	0.46	.16	.11
1109	2064	0.00	46.2	9	3.98	3.00	1.25	1.27	.15	.03
1124	1115	-.01	29.0	6	4.00	3.00	1.39	1.42	.22	.03
1214	3944	0.01	15.6	15	2.87	1.98	0.46	0.57	.28	.07
1227	4332	0.01	37.3	19	2.23	1.61	0.29	0.25	.12	.04
1230	3414	0.00	40.8	12	4.00	3.00	0.96	1.00	.20	.04
1320	1634	0.00	18.4	14	1.63	0.85	-1.09	-1.28	.12	.07
1322	4005	0.01	87.2	19	2.14	1.58	0.01	0.20	.14	.11
1328	3522	0.01	85.4	19	1.94	1.60	0.00	0.07	.19	.04
1329	1692	0.00	28.5	14	1.39	0.86	-0.71	-0.88	.23	.03
1414	1438	0.00	12.0	6	4.00	2.26	1.29	1.40	.23	.07
1509	2570	0.00	17.7	10	3.56	2.70	1.08	1.08	.13	.04
1526	3096	0.00	8.8	11	2.55	2.44	0.92	0.89	.07	.03

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-8

RESULTS OF ITEM RECALIBRATION FOR SUBTEST MC

Item	N	ASD*	ChiSq	DF	Item parameters A PP	A CAT	B PP	B CAT	C	AAD**
804	3624	0.00	15.0	8	1.54	1.84	0.16	0.56	.26	.19
807	1277	0.00	17.1	10	0.81	1.25	-0.49	-0.14	.10	.15
808	5173	0.00	12.3	18	1.59	1.20	0.08	-0.16	.16	.08
810	5349	0.00	85.4	15	3.00	1.83	0.02	-0.10	.20	.06
812	5151	0.00	22.5	19	1.72	1.52	-0.05	0.02	.20	.04
813	1146	0.00	3.7	3	1.08	1.36	0.15	-0.02	.17	.07
814	4919	0.00	16.4	13	3.00	1.97	0.32	0.11	.27	.10
815	1291	0.00	6.2	6	1.01	1.24	-0.28	-0.11	.25	.07
818	4719	0.00	27.0	12	2.54	1.76	0.38	0.73	.27	.17
821	1336	0.00	11.9	8	1.38	1.69	0.97	1.20	.14	.13
823	2720	0.00	9.4	10	1.71	1.77	0.72	1.26	.23	.24
824	1844	0.00	7.3	8	1.41	1.51	0.73	0.58	.19	.07
902	2556	0.00	16.8	15	1.09	1.41	-0.72	-0.42	.18	.09
909	5540	0.00	27.9	17	1.74	1.37	0.04	0.12	.15	.05
910	4037	0.00	16.4	12	1.37	1.48	-0.02	-0.01	.17	.01
912	2754	0.00	17.1	9	1.12	1.53	-0.05	-0.37	.16	.15
914	3193	0.00	9.1	9	1.27	1.59	0.09	0.02	.17	.04
916	4046	0.00	11.1	14	3.00	1.67	0.61	0.34	.34	.13
921	2763	0.00	9.1	12	1.75	1.42	0.81	0.76	.18	.04
924	1197	0.00	9.5	11	3.00	3.00	1.45	1.58	.28	.04
925	1292	-.01	19.4	11	3.00	1.88	1.46	1.26	.26	.12
1001	1839	0.01	8.9	13	0.92	1.39	-0.79	-0.77	.20	.04
1004	1241	0.00	7.0	7	1.03	1.33	-0.40	-0.42	.31	.02
1005	3041	0.00	11.4	9	1.30	0.95	-0.01	-0.16	.20	.05
1007	2621	0.00	27.4	14	1.09	1.34	-0.57	-0.25	.21	.10
1008	1320	0.00	12.9	12	0.86	0.94	-0.79	-0.34	.20	.13
1013	5295	0.00	23.7	16	2.11	1.32	0.29	0.05	.18	.09
1019	3850	0.00	3.9	9	1.48	1.51	0.49	0.57	.19	.04
1020	3387	0.00	12.9	13	3.00	1.55	0.81	0.51	.23	.18
1021	4119	0.00	14.8	15	3.00	1.60	0.76	0.80	.18	.07
1025	1800	0.00	14.9	11	1.68	0.73	1.05	1.49	.17	.06

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

TABLE A-9

RESULTS OF ITEM RECALIBRATION FOR SUBTEST EI

Item	N	ASD*	ChiSq	DF	A PP	A CAT	B PP	B CAT	C	AAD**
					Item parameters					
712	2424	0.00	25.7	10	1.38	1.41	-0.74	-0.72	.38	.01
723	2516	0.00	26.1	10	1.41	1.37	-0.29	-0.33	.32	.01
728	3494	0.00	20.5	22	1.58	1.04	-0.02	-0.29	.31	.07
733	3957	0.00	47.5	19	1.50	1.62	-0.42	0.00	.24	.17
734	4670	0.00	41.5	25	1.71	1.06	-0.02	-0.03	.22	.05
737	2399	0.00	27.5	8	1.49	1.06	-0.13	-0.17	.33	.02
744	1021	0.00	20.7	7	2.08	1.66	1.24	1.49	.23	.08
747	1046	0.00	5.1	4	1.37	0.70	0.11	0.45	.30	.08
750	4345	0.00	58.9	25	1.74	1.39	0.28	0.38	.18	.04
752	3131	0.00	47.0	21	1.61	1.15	0.51	0.19	.18	.12
761	1722	0.00	12.8	6	1.38	1.39	0.14	0.57	.28	.16
762	1312	0.00	22.5	5	1.48	1.21	0.30	0.42	.33	.05
763	1678	0.00	12.0	10	1.41	1.83	-1.33	-1.36	.31	.04
766	2657	0.00	34.4	16	1.73	1.48	0.61	0.50	.21	.05
804	2714	0.00	24.3	15	1.39	1.32	-0.96	-0.89	.34	.02
806	3307	0.00	23.4	16	1.29	0.98	-0.70	-1.12	.25	.09
827	1039	0.00	1.3	2	1.14	2.04	-0.62	-0.56	.31	.03
831	4301	0.00	56.8	24	1.50	1.22	-0.28	0.05	.25	.11
832	2536	0.00	12.5	9	1.07	1.19	-0.64	-0.63	.14	.01
833	4757	0.01	80.3	25	1.72	1.28	-0.41	-0.27	.20	.06
837	1448	0.00	5.5	5	1.28	1.46	-0.25	-0.66	.31	.15
840	2383	0.00	22.1	9	1.39	1.02	0.20	-0.25	.24	.14
842	3705	0.00	22.0	22	1.61	1.14	0.54	0.65	.08	.05
845	1947	0.00	17.4	9	1.48	1.84	0.37	0.60	.28	.10
855	1134	0.00	18.8	6	1.38	1.65	0.84	0.58	.18	.13
856	2179	0.00	33.0	17	2.59	1.75	0.95	0.78	.22	.10
869	1941	0.00	7.7	13	1.38	1.14	-1.52	-1.66	.07	.01
874	1722	0.00	13.3	9	1.25	1.17	-1.30	-1.38	.24	.01
875	2085	0.00	26.8	13	1.25	1.13	-1.27	-1.92	.19	.13
881	2230	0.00	24.0	8	1.21	1.09	-0.88	-0.69	.26	.07
889	1225	0.00	17.6	13	3.00	1.86	1.33	1.15	.26	.12
7103	1971	0.00	26.8	17	2.03	1.54	0.91	0.91	.24	.03
7105	1598	0.00	28.8	14	1.80	1.13	1.21	1.41	.08	.05

* ASD = Average (observed CAT proportion - CAT IRC)
** AAD = Average absolute (CAT IRC - PP IRC)

(12)

CRM 86-189 / August 1986

AD-A175 507

RESEARCH MEMORANDUM

DETERMINING THE SENSITIVITY OF CAT-ASVAB SCORES TO CHANGES IN ITEM RESPONSE CURVES WITH THE MEDIUM OF ADMINISTRATION

D. R. Divgi

A Division of CNA Hudson Institute
CENTER·FOR·NAVAL·ANALYSES

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Work conducted under contract N00014-83-C-0725.

A Division of CNA Hudson Institute
CENTER·FOR·NAVAL·ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

3 September 1986

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 86-189

Encl: (1) CNA Research Memorandum 86-189, "Determining the Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves With the Medium of Administration," by D. R. Divgi, Aug 1986

1. Enclosure (1) is forwarded as a matter of possible interest.

2. The Department of Defense may implement a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) in the near future. Analysis of an experimental version has shown that characteristics of items change from paper-pencil to CAT administration. This research memorandum examines the effects of these changes on the CAT-ASVAB scores of individual examinees.

William H. Sims
Director
Marine Corps Manpower
and Training Program

Distribution List:
Reverse Page

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release; Distribution Unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 86-189	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Center for Naval Analyses 6b. OFFICE SYMBOL (If applicable) CNA	7a. NAME OF MONITORING ORGANIZATION Commandant of the Marine Corps (Code RDS)
6c. ADDRESS (City, State, and ZIP Code) 4401 Ford Avenue Alexandria, Virginia 22302-0268	7b. ADDRESS (City, State, and ZIP Code) Headquarters, Marine Corps Washington, D.C. 20380
8a. NAME OF FUNDING / ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable) ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-83-C-0725
8c. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 65153M PROJECT NO. C0031 TASK NO WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Determining the Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves With the Medium of Administration

12. PERSONAL AUTHOR(S)
D. R. Divgi

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM TO	August 1986	36

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	10	
12	01	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Administration, Aptitude tests, ASVAB (Armed Services Vocational Aptitude Battery), Bayes Theorem, Calibration, CAT (Computerized Adaptive Testing), Chi-Square test, Equations, Estimates, IRC (Item Response Curve), Mathematical Analysis, Mean, Mental ability, PP (paper-pencil), Recruits, Test methods, Test scores

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Within a few years the Department of Defense may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). In CAT, each test item is characterized by an item response curve (IRC), which describes how the probability of correctly answering the item increases with ability. A recent study conducted by the Center for Naval Analyses found that IRCs of many items in the experimental CAT item pool for the ASVAB changed substantially from paper-pencil to CAT administration. This research memorandum examines the effects of these changes on scores of individual examinees.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt.Col. G. W. Russell	22b. TELEPHONE (Include Area Code) (202) 694-3491 22c. OFFICE SYMBOL RDS-40

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE